تن سازان، امیر؛ مهدوی، محمدامین (۱۳۹۶). استخراج فرادادههای متنی از مقالههای علمی به زبان فارسی با مدل آماری CRF. *پژوهش‌های نظری و کاربردی در علم اطلاعات و دانش‌شناسی*، ۷ (۱)، ۳۰۴- ۳۲۱.

استخراج فرادادههای متنی از مقالههای علمی به زبان فارسی با مدل آماری CRF

امیر تن سازان، دانشجوی کارشناسی ارشد مهندسی کامپیوتر- گرایش نرم افزار، دانشگاه بین المللی امام خمینی (ره) قزوین، Ta.40.amir@gmail.com

محمد امین مهدوی، استادیار دانشکده فنی و مهندسی، دانشگاه بین المللی امام خمینی (ره) قزوین، mahdavi@eng.ikiu.ac.ir

تاریخ دریافت: ۹۵/۶/۸ تاریخ پذیرش: ۹۵/۸/۱۸

چکیده:

مقدمه: *استخراج فرادادههای متنی از مقالههای علمی به شکل دستی کار زمان‌بر و پرهزینه‌ای است. وجود تنوع در قالب‌های ساختاری مقالات علمی نیز به پیچیدگی مسئله می‌افزاید. بنابراین، استخراج خودکار فرادادههای متنی از مقالههای علمی به عنوان یک مسئله مطرح است و از الگوریتم‌های مختلفی می‌توان برای استخراج فرادادهها استفاده کرد. هدف این مقاله ارائه‌ی یک چارچوب برای استخراج فرادادههای متنی از مقالههای علمی به زبان فارسی است. در این پژوهش از مدل آماری سی آر اف برای استخراج فرادادهها استفاده شده است.*

روش‌شناسی: *این مقاله یک پژوهش کاربردی است. در این مقاله با مطالعات کتابخانه‌ای و آزمایش سعی شده است یک چارچوب برای استخراج فرادادهها ارائه شود. چارچوب ارائه شده شامل شناسایی سرآیند و مراجع انگلیسی و فارسی مقاله است. از مدل آماری سی آر اِف برای استخراج فرادادهها از سرآیند و مراجع فارسی و انگلیسی استفاده شده است. با تعریف ویژگی‌های مختلف این مدل آماری قابل تغییر است. آزمایش این روش بروی صد مقاله از مجلات علمی- پژوهشی ایران درصد موفقیت آن را نشان می‌دهد. مدل آماری سی آر اِف در برچسب‌زنی متن نسبت به مدل‌های آماری دیگر مانند مدل مخفی مارکوف دقت بالاتری را ارائه می‌دهد. از سوی دیگر این مدل بر مبنای آمار و ریاضی برچسب‌زنی را انجام می‌دهد. استخراج فرادادهها از مقالات با*

دوفصلنامه| (علمی پژوهشی)
پژوهش‌های نظری و کاربردی در علم اطلاعات و دانش شناسی

شاپا (آنلاین): ۲۵۳۸-۴۱۱۲

http://infosci.um.ac.ir

سال ۷ (شماره ۱)
بهار و تابستان ۱۳۹۶

DOI: 10.22067/58517

قالب‌های مختلف به کمک آمار نسبت به روش‌های مبتنی بر قانون نتایج بهتری را به دنبال دارد. بنابراین استفاده از مدل آماری سی آر اِف برای حل این مسئله مناسب است.

یافته‌ها: *برای ارزیابی روش پیشنهاد شده از معیار اِف استفاده شده است. مقدار معیار اِف در این پژوهش برای هر توکن متنی محاسبه شده است. مقدار معیار اِف به شکل میانگین برای فراداده‌های سرآیند، فراداده‌های مراجع فارسی و فراداده‌های مراجع انگلیسی به ترتیب ۹۶/۸۹ درصد، ۹۳/۸۷ درصد و ۹۴/۷۵ درصد است. نتایج این پژوهش با سه پژوهش مشابه در زبان انگلیسی مقایسه شده است. مقایسه میانگین نتایج به‌دست آمده نشان می‌دهد در فراداده‌های سرآیند نتایج پژوهش این مقاله بهتر از دو پژوهش انجام شده در زبان انگلیسی است. نتایج استخراج فراداده نویسنده در سرآیند در پژوهش‌های زبان انگلیسی بهتر است. برای فراداده چکیده در پژوهش زبان فارسی، نتایج بهتری به‌دست آمده است. مقایسه میانگین نتایج استخراج فراداده‌های مراجع، نشان می‌دهد پژوهش‌های زبان انگلیسی دقت بالاتری ارائه داده‌اند. نتایج استخراج فراداده مؤسسه در مراجع فارسی نسبت به فراداده‌های دیگر ضعیف‌تر است.*

بحث و نتیجه‌گیری: *بررسی نتایج بدست آمده نشان می‌دهد که عملکرد مدل آماری سی آر اِف برای استخراج فراداده‌ها خوب است. بیشترین دقت برای فراداده چکیده با معیار اِف برابر ۹۹/۶ درصد است. این فراداده تعداد توکن بسیار بیشتری نسبت به بقیه فراداده‌ها دارد. دقت فراداده مؤسسه با معیار اِف برابر ۸۰/۹۵ درصد کمتر از بقیه است. دو دلیل در کاهش دقت موثر است. تعداد این فراداده در پیکره متون نسبت به فراداده‌های دیگر کمتر است. علاوه بر این کلمات نحوی که در این فراداده به‌کار می رود، تنوع بیشتری دارد. در مراجع فارسی اسامی شهرها در فراداده‌های مکان و مؤسسه به‌کار می‌رود. این مسئله باعث می‌شود در برخی از موارد فراداده‌های مکان و مؤسسه به اشتباه تشخیص داده شوند. در زبان فارسی کلماتی که به شکل مشترک در فراداده‌های مختلف به‌کار می‌روند نسبت به زبان انگلیسی بیشتر است. برای مثال بسیاری از اسامی ایرانی که برای نام افراد به‌کار می‌رود با معانی دیگر در فراداده‌های دیگر استفاده می‌شود. این مسئله ممکن است باعث بروز خطا شود. اکثر خطاهای به‌وجود آمده در استخراج فراداده‌ها مربوط به توکن‌هایی است که در مرز دو فراداده قرار دارند. تبدیل مقالات علمی فارسی با فرمت پی دی اِف به فرمت متن در موارد زیادی با مشکل رو به رو است و از محدودیت‌های این پژوهش به شمار می‌آید. در این پژوهش مجموعه‌ای از صد مقاله علمی استفاده شد. افزایش تعداد مقاله‌های علمی و تنوع بیشتر مقالات برای آزمایش می‌تواند در نتیجه‌ی بدست آمده تاثیر مثبتی داشته باشد. مجموعه‌ای از ویژگی‌های متنی در الگوریتم‌های برچسب‌زنی سی آر اِف استفاده می‌شود. تغییر در این ویژگی‌ها می‌تواند موجب بهینه‌سازی روش شود.*

کلیدواژه‌ها: *استخراج فراداده‌های متنی، مقاله‌های علمی، پردازش زبان فارسی، الگوریتم CRF.*

مقدمه

مقاله‌های علمی نقش مهمی در دنیای پژوهش دارند. آخرین یافته‌های علمی در حوزه‌های مختلف در مقاله‌ها منتشر می‌شود. دسترسی به یافته‌های جدید و پژوهش‌های روز از ملزومات تحقیق در یک زمینه علمی است. دسترسی به پژوهش‌های جدید و مرتبط، به جستجو نیاز دارد. حجم انبوه مقاله‌ها و نبود فراداده‌های متنی مقالات، امکان جستجوی مقاله‌های علمی را دشوار می‌کند (Tkaczyk et al., 2015).

امروزه کتابخانه‌های دیجیتال و وب سایت‌های مجلات علمی و موتورهای جستجو، دسترسی و جستجوی هوشمند مقاله‌های علمی منتشر شده را برای پژوهشگران فراهم می‌کنند. این کتابخانه‌ها برای نمایه‌سازی مقاله‌های علمی به فراداده‌های متنی[1] مقاله‌ها نیاز دارند (Tkaczyk et al., 2015). اطلاعاتی مانند نویسنده، عنوان، تاریخ، چکیده و نام انتشارات، فراداده‌های مقاله‌های علمی هستند. علاوه بر این، کتابخانه‌های دیجیتال برای نمایه‌سازی مراجع مقالات و نمایش مقاله‌های مرتبط به یک جستجو و محاسبه شاخص ارجاع[2] به فراداده‌های مراجع نیاز دارند. نام نویسندگان، عنوان، منبع، تاریخ و مؤسسه از جمله فراداده‌های مراجع در مقاله‌های علمی هستند (Guo and Jin, 2011b).

فراداده‌های مقاله را می‌توان در سه دسته قرار داد. دسته اول، فراداده‌های سرآیند[3] مقاله که شامل اطلاعاتی مانند عنوان، چکیده و نویسندگان است. دسته دوم، فراداده‌های مراجع که اطلاعاتی مانند نویسنده، عنوان، تاریخ و منبع در هر مرجع است. دسته سوم، فراداده‌های بدنه مقاله است. عناوین اصلی و فرعی مقاله از جمله فراداده‌های بدنه مقاله است. اکثر روش‌ها برای فراداده‌های سرآیند یا مراجع ارائه شده‌اند.

استخراج فراداده‌های متنی به شکل دستی کار هزینه‌بری است. از این‌رو ایجاد روش‌هایی برای استخراج خودکار فراداده‌های متنی مقالات ضروری است. این کار به‌دلیل تنوع فرمت مقالات و سبک‌های مختلف مراجع آنها چالش محسوب می‌شود (Guo and Jin, 2011a).

در این مقاله چارچوبی برای استخراج فراداده‌های متنی مقاله‌های علمی به زبان فارسی ارائه می‌شود. در بخش اول کارهایی که در گذشته در این زمینه در زبان‌های دیگر انجام شده است، مرور می‌شود. در بخش دوم الگوریتم برچسب‌زنی سی آر اِف[4] معرفی می‌شود. در بخش سوم روش پیشنهادی برای مقاله‌های فارسی بیان می‌شود. در بخش چهارم پیکره متونی که تست روش بر روی آنها انجام شده است و نتایج به‌دست آمده از تست روش بروی پیکره متون ارائه می‌شود.

مروری بر کارهای گذشته

تاکنون روش‌های مختلفی برای استخراج فراداده‌های متنی در زبان انگلیسی ارائه شده است. این‌روش‌ها را می‌توان در دو دسته قرار داد. دسته اول روش‌های مبتنی بر قاعده و الگو و دسته دوم

1. metadata
2. Citation index
3. header
4. Conditional Random Field

روش‌های مبتنی بر یادگیری (Tkaczyk et al., 2015).

در روش‌های مبتنی بر قاعده[۱] قوانینی تعریف می‌شود که با استفاده از آنها فراداده‌های متنی مقاله استخراج می‌شوند. گو و جین[۲] (۲۰۱۱a) چارچوبی را برای استخراج فراداده‌های سرآیند ارائه داده است. در این چارچوب برای استخراج فراداده‌ها، مجموعه قوانین مرتبط با ویژگی‌های فرمت و محتوای متن مقاله تعریف شده است. تکه‌های متنی مقاله و ویژگی‌های فرمت از تبدیل پی دی اف[۳] به فرمت ایکس اِم اِل[۴] فراهم شده است. جیوفریدا، شِک و یانگ[۵] (۲۰۰۰) روشی برای استخراج فراداده‌های سرآیند مقاله‌ها با فرمت پست اسکریپت[۶] ارائه کرده است. در این‌روش از ویژگی‌های ظاهری و بصری مقاله استفاده می‌شود. این ویژگی‌ها توسط ابزار پی اِس تو تکست[۷] فراهم می‌شود. بیل[۸] و همکاران (۲۰۱۰) روشی مبتنی مبتنی بر قانون برای استخراج عنوان ارائه داده است که در آن پس از تبدیل فرمت پی دی اِف به فرمت ایکس اِم اِل از اندازه فونت و موقعیت مکانی متن برای شناسایی عنوان استفاده شده است. گو و جین (۲۰۱۱b) برای استخراج فراداده‌های مراجع مقاله روشی بر مبنای تطبیق الگو ارائه داد. در این‌روش براساس سبک‌های مختلف مرجع نویسی، مجموعه‌ای از الگوها تعریف شده و با تطبیق الگو فراداده‌های هر مرجع استخراج می‌شود.

روش‌های مبتنی بر یادگیری برای استخراج فراداده‌های متنی در قالب طبقه‌بندی[۹] و برچسب‌زنی[۱۰] دنباله قرار می‌گیرند. بخش‌هایی که برای طبقه‌بندی و برچسب‌زنی مورد استفاده قرار می‌گیرد، بلاک، سطر و کلمات مقاله است (Tkaczyk et al., 2015). کواسویچ[۱۱] و همکارانش (۲۰۱۱) روشی را براساس الگوریتم اِس وی اِم[۱۲] برای استخراج فراداده‌های سرآیند ارائه دادند. در این‌روش سطرهای مقاله طبقه‌بندی شده‌اند و ویژگی‌های مرتبط با کلمه و سطر برای طبقه‌بندی تعریف و استفاده شده است. از ابزار پی دی اِف تو اِچ تی اِم اِل[۱۳] برای استخراج ویژگی‌های فرمت استفاده شده است. سیمور، مَکالوم و روزنفلد[۱۴]

1. Rule based
2. Guo and Jin
3. Portable Document Format
4. XML
5. Giuffrida, Sheck and Yang
6. PostScript
7. PsToText
8. Beel
9. Classification
10. labeling
11. Kovacevic
12. Support Vector Machine
13. Pdftohtml
14. Seymore & McCallum & Rosenfeld

(۱۹۹۹) از الگوریتم مدل مخفی مارکوف برای برچسب‌زنی کلمات سرآیند مقاله و استخراج فراداده‌های متنی استفاده کرد. پِنگ و مَک‌الوم[۱] (۲۰۰۶) ایده به‌کارگیری الگوریتم سی آر اِف را برای استخراج فراداده‌های سرآیند و فراداده‌های مراجع را مورد بررسی قرار داد. در این‌روش، دنباله‌ای از کلمات سرآیند و دنباله‌ای از کلمات هر مرجع برچسب‌زنی شده‌اند. هان[۲] و همکارانش (۲۰۰۳) روشی براساس الگوریتم اِس وی اِم ارائه دادند. در این‌روش از طبقه‌بندی دو مرحله‌ای برای سطرهای مقاله استفاده می‌شود. برای استخراج ویژگی‌های طبقه‌بندی از خوشه‌بندی کلمات مقاله استفاده شده است. کانسیل، گیلز و کان[۳] (۲۰۰۸) از الگوریتم سی آر اِف برای استخراج فراداده‌های مراجع استفاده کرد. شناسایی و تفکیک مراجع قبل از پردازش توسط قواعد ساده انجام شده است. ژانگ[۴] و همکارانش (۲۰۱۱) از الگوریتم اِس وی اِم ساختاریافته[۵] برای پارس مراجع مقاله‌های علمی استفاده کرد. در این‌روش علاوه بر ویژگی‌های توکن از ویژگی‌های توکن‌های همسایه در برچسب‌زنی هر مرجع استفاده شده است. هتزنر[۶] (۲۰۰۸) روشی را برای استخراج فراداده‌های مراجع براساس مدل مخفی مارکوف ارائه داد.

تکاسزیک و همکارانش (۲۰۱۵) چارچوبی را برای استخراج فراداده‌های سرآیند و مراجع مقاله ارائه داد که کل متن مقاله را پردازش می‌کند. در این چارچوب از کتابخانه آی تکست[۷] و الگوریتم داک استرام[۸] برای قطعه‌بندی متن مقاله استفاده شده است. الگوریتم اِس وی اِم برای طبقه‌بندی بلوک‌های متنی مقاله و الگوریتم سی آر اِف برای استخراج فراداده‌های مراجع استفاده شده است. تفکیک مراجع نیز با استفاده از خوشه‌بندی نزدیک‌ترین همسایگی انجام شده است. کاندیاس[۹] (۲۰۱۱) چارچوبی را ارائه داد که در آن علاوه بر فراداده‌های سرآیند و مراجع، بدنه و عناوین بخش‌های مقاله استخراج می‌شود. از الگوریتم سی آر اِف برای پردازش استفاده شده است. ابزار پی دی اِف تو اِچ تی اِم اِل برای استخراج ویژگی‌های فرمت در این‌روش استفاده شده است.

ابزارهای مختلفی در زبان انگلیسی برای استخراج فراداده‌ها از مقاله‌های علمی وجود دارد. برای مثال سی بی تو بیب[۱۰] نرم‌افزاری رایگان برای استخراج داده‌های کتاب‌شناسی از مقاله‌های پی دی اِف و

1. Peng and McCallum
2. Han
3. Councill, Giles and Kan
4. Zhang
5. Structural SVM
6. Hetzner
7. iText
8. Docstrum
9. Candeias
10. Cb2bib

صفحات وب است. در این نرم‌افزار مجموعه‌ای از الگوهای از پیش تعریف شده برای استخراج فرادادهها استفاده می‌شود. فراداده‌های استخراج شده با فرمت بیب تکس[1] ذخیره می‌شوند. امکان مدیریت فایل‌های کتاب‌شناسی و مقالات در این نرم‌افزار فراهم شده است (Cb2bib Overview, 2016). پارس سایت[2] ابزار ابزار دیگری است که از الگوریتم سی آر اِف برای استخراج فراداده‌های متنی از مراجع مقاله‌های علمی و استخراج ساختار مقاله‌های انگلیسی استفاده می‌کند. امکان ذخیره‌سازی فراداده‌های استخراج شده با فرمت‌های مختلف مانند بیب تکس[3] و ایکس ام ال در این ابزار فراهم شده است (Parscit, 2016).

مدل آماری سی آر اِف

مدل سی آر اِف شکلی از مدل گرافی هدایت نشده[4] است که توزیع خطی- لگاریتمی روی دنباله برچسب‌ها را براساس دنباله مشاهدات داده شده تعریف می‌کند. احتمال شرطی دنباله برچسب‌های $Y = y_1 \ldots y_i$ و دنباله مشاهدات $X = x_1 \ldots x_j$ به شکل زیر است (Wallach, 2004):

$$P(y|x, \lambda) = \frac{1}{Z(x)} \exp\left(\sum_j \sum_{i=1}^{n} \lambda_j f_j\left(y_{i-1}, y_i, x, i\right)\right) \quad (۱)$$

در این رابطه $Z(x)$ ثابت نرمال‌سازی[5] است. $f_j\left(y_{i-1}, y_i, x, i\right)$ تابع ویژگی است. توابع ویژگی برای برای بیان خصوصیات داده‌ها، مجموعه‌ای از ویژگی‌های مشاهدات را تعریف می‌کند. تابع ویژگی[6] به دو شکل تابع حالت[7] و تابع انتقال[8] نمایش داده می‌شود. برای مثال یک تابع انتقال به شکل زیر تعریف می‌شود:

$$t_j\left(y_{i-1}, y_i, x, i\right) = \begin{cases} b(x, i) & \text{if } y_{i-1} = \text{abstract and } y_i = \text{keyword} \\ 0 & \text{otherwise} \end{cases} \quad (۲)$$

$b(x,i)$ یک ویژگی دو یا چند مقداری است. برای مثال:

$$b(x,i) = \begin{cases} 1 & \text{if the observation at position i is a word "abstract"} \\ 0 & \text{otherwise} \end{cases} \quad (۳)$$

تابع حالت فقط برچسب فعلی را در نظر می‌گیرد. $\lambda = \lambda_1 \ldots \lambda_j$ پارامترهای توابع ویژگی هستند که از داده‌های یادگیری تخمین زده می‌شوند.

با مدل داده شده در معادله (۱) محتمل‌ترین دنباله برچسب‌ها برای دنباله مشاهدات به شکل زیر است (Lafferty, McCallum and Pereira, 2001):

1. Bibtex
2. ParsCit
3. Bibtex
4. Undirected graphical model
5. Normalization constant
6. Feature function
7. State function
8. Transition function

$$Y^{*} = \underset{y}{\arg\operatorname{Max}}\, P(y|x) \qquad (۴)$$

این معادله با کمک برنامه‌نویسی دینامیک و الگوریتم ویتربی[1] محاسبه می‌شود.

تخمین پارامترها در الگوریتم سی آر اِف با روش بیشترین درست‌نمایی[2] انجام می‌شود. در صورتی‌که داده‌های یادگیری به شکل $\{(x_k, y_k): k = 1,\ldots M\}$ باشد، حاصل ضرب معادله (۱) روی داده‌های یادگیری به‌عنوان تابعی از پارامترهای λ به‌عنوان تابع درست‌نمایی شناخته می‌شود و با $P(\{y^k\}|\{x^k\}, \lambda)$ نمایش داده می‌شود. روش بیشترین درست‌نمایی، پارامترهای λ را به گونه‌ای انتخاب می‌کند که الگوریتم درست‌نمایی با عنوان لگاریتم درست‌نمایی[3] بیشترین مقدار را داشته باشد. برای الگوریتم سی آر اِف مقدار لگاریتم درست‌نمایی به شکل زیر تعریف می‌شود (Wallach, 2004):

$$L(\lambda) = \sum_k \left(\sum_j \sum_{i=1}^{n} \lambda_j f_j\left(y_{i-1}, y_I, x, i\right) - \log\left(Z(x_k)\right) \right) \qquad (۵)$$

ماکزیمم کردن رابطه (۵) معادله تساوی (۶) را به‌وجود می‌آورد. در این رابطه تعداد تجربی هر ویژگی با تعداد مورد انتظار آن در مدل $P(y|x)$ تطبیق می‌یابد.

$$\sum_k \sum_i f_j\left(y_{i-1}, y_i, x_k, i\right) = \sum_k \sum_{y'} p(y'|x_k) \sum_i f_j(y'_{i-1}, y'_i, x_k, i) \qquad (۶)$$

برای حل رابطه بالا و تخمین پارامترها، الگوریتم مقیاس‌گذاری تکراری[4] ارائه شده است. این الگوریتم سرعت کمی دارد. اثبات شده است که الگوریتم L-BFGS[5] با سرعت بیشتری به ماکزیمم دست دست می‌یابد. بنابراین برای تخمین پارامترها از این الگوریتم استفاده می‌شود (Peng and McCallum, 2006).

روش استخراج فراداده‌ها

در اینجا، مقالات علمی فارسی با فرمت متن به‌عنوان ورودی در نظر گرفته شده است؛ از این‌رو پردازش بر روی متن انجام می‌گیرد. به عبارتی، متن باید توسط ماشین قابل خواندن باشد. چارچوبی که برای استخراج فراداده‌های متنی از مقاله‌های علمی پیشنهاد شده است، به سه مرحله قابل تقسیم است. در مرحله اول بخش سرآیند و بخش مراجع در مقاله شناسایی می‌شود. در مرحله دوم مراجع فارسی و انگلیسی از بخش مراجع شناسایی و تفکیک می‌شوند. پس از این کار مراجع مختلف مقاله شناسایی و از یکدیگر تفکیک می‌شوند. در مرحله سوم، سرآیند و مراجع مختلف فارسی و انگلیسی توسط سه

---

1. Viterbi
2. Maximum likelihood
3. log-likelihood
4. iterative Scaling
5. limited memory broyden fletcher goldfarb shanno

برچسب‌زن مجزای سی آر اِف پردازش شده و فراداده‌های سرآیند، مراجع فارسی و مراجع انگلیسی استخراج می‌شود. شکل ۱ نمای کلی این چارچوب را نشان می‌دهد.

شناسایی سرآیند و مراجع

فراداده‌هایی که در روش پیشنهادی استخراج می‌شود، در سرآیند یا بخش مراجع مقاله قرار دارند؛ بنابراین، این دو بخش قبل از پردازش باید شناسایی شوند. برای این کار از ویژگی‌های متنی استفاده می‌شود. شروع متن مقاله ابتدای سرآیند مقاله است. انتهای سرآیند، شروع بخش مقدمه در مقاله است؛ بنابراین با کلمه کلیدی «مقدمه» این بخش شناسایی می‌شود. در برخی از مقاله‌های فارسی قبل از بخش مقدمه، عنوان و چکیده انگلیسی قرار گرفته است. در روش پیشنهادی، بخش انگلیسی جزئی از سرآیند نیست؛ بنابراین در این حالت از قاعده دیگری برای شناسایی سرآیند استفاده می‌شود. در این حالت با پیمایش متن مقاله و رسیدن به کلمه "abstract" قبل از مقدمه مقاله، پیمایش متن متوقف می‌شود و تا رسیدن به کلمات فارسی عقب‌گرد صورت می‌گیرد.

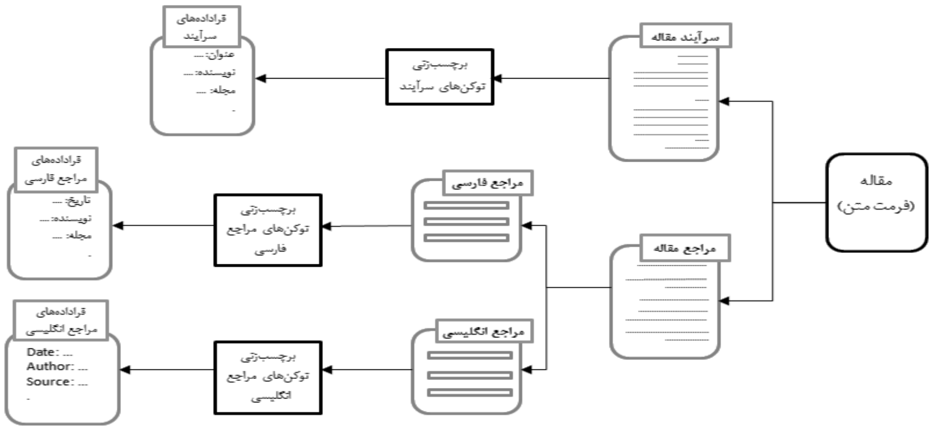
شناسایی ابتدای بخش مراجع با استفاده از کلمات کلیدی مانند «منابع» و «مراجع» و "References" انجام می‌شود. در صورتی که پس از مراجع، متن دیگری در مقاله نباشد، انتهای مقاله، انتهای مراجع نیز خواهد بود. در غیر این صورت چند حالت ممکن است رخ دهد. در حالت اول پس از مراجع بخش دیگری قرار دارد. در این حالت به کمک کلمات کلیدی عناوین بخش‌های مقاله و طول سطرهای مقاله، انتهای مراجع شناسایی می‌شود. در حالت دوم در برخی از مقاله‌های فارسی، عنوان و چکیده انگلیسی پس از بخش مراجع قرار گرفته است. در این حالت پس از پیمایش سطرهای مقاله از شروع بخش مراجع با رسیدن به سطری شامل کلمه "abstract" عقب‌گرد انجام می‌شود. این عقب‌گرد تا رسیدن به سطری که با علامت نقطه پایان یافته است ادامه می‌یابد. علت این کار این است که هر مرجعی در بخش مراجع با نقطه پایان یافته است.

مراحل تفکیک مراجع

بسیاری از مقاله‌های فارسی شامل مراجع انگلیسی و فارسی هستند. برای استخراج فراداده‌ها، این مراجع باید از یکدیگر تفکیک شوند. در مرحله اول مراجع فارسی و انگلیسی براساس الگو شناسایی و از یکدیگر تفکیک می‌شوند. برای این کار از تفاوت کاراکترهای انگلیسی و فارسی استفاده می‌شود. با این قاعده، در دو حالت امکان خطا وجود دارد. حالت اول زمانی است که سطری از بخش مراجع شامل کاراکترهای انگلیسی و فارسی باشد. در این حالت این سطر جزء مراجع فارسی است. زیرا گاهی در

مراجع فارسی کلمات انگلیسی به‌کار می‌رود. حالت دوم زمانی است که در مرجع فارسی یک سطر از کلمات انگلیسی وجود داشته باشد. در این حالت از تفاوت طول یک سطر با سطر قبلی‌اش در مراجع و کاراکترهای شروع مراجع مانند اعداد برای قرار دادن یک سطر در مراجع انگلیسی یا فارسی استفاده می‌شود.

پس از تفکیک مراجع انگلیسی و فارسی، مراجع مختلف از یکدیگر جدا می‌شوند. این کار با شناسایی سطرهای شروع یا پایان هر مرجع امکان‌پذیر است. مراجع مقاله‌های مختلف در دو دسته قرار می‌گیرند. دسته اول مراجعی هستند که با نشانه‌هایی مانند «[۱]» و «-» شروع می‌شوند. دسته دوم بدون نشانه‌اند. در دسته اول، نشانه‌های شروع برای تشخیص سطر اول هر مرجع به‌کار گرفته می‌شود. مقاله‌های فارسی به دو شکل یک ستونه و دو ستونه نوشته می‌شوند. برای دسته دوم، مراجع با توجه به طول سطر، یکی از دو حالت یک ستونه و دوستونه را خواهند داشت. در حالت یک ستونه تعداد سطر هر مرجع کمتر است و برخی از مراجع یک سطری هستند. در حالت دو ستونه تعداد سطرهای هر مرجع بیشتر از یک است.



شکل ۱. استخراج فرادادهها از مقاله‌های علمی فارسی

برای حالت دو ستونه از دو ویژگی برای تشخیص سطر انتهایی هر مرجع استفاده می‌شود. ویژگی اول این است که هر مرجع با علامت نقطه پایان می‌پذیرد. ویژگی دوم تفاوت طول یک سطر با سطر قبلی‌اش در مراجع است. برای حالت یک ستونه نیز دو ویژگی برای شناسایی سطر پایانی استفاده می‌شود. ویژگی اول پایان یافتن هر مرجع با علامت نقطه و ویژگی دوم اینکه سطر ابتدایی هر مرجع شامل نام نویسنده یا حالت مخفف نام است.

مراحل استخراج فراداده‌ها

سرآیند مقاله و هر یک از مراجع فارسی و انگلیسی که در مرحله قبل شناسایی شده است، در این مرحله ابتدا به مجموعه‌ای از توکن‌ها تجزیه می‌شوند و سه مجموعه توکن تشکیل می‌شود. تجزیه تنها با فضای خالی انجام می‌شود. علائم نگارشی به کلمه قبل از خود در بسیاری از موارد متصل هستند. کلماتی مانند «ها» که در زبان فارسی به شکل پیوسته و جدا نوشته می‌شوند، مشکلی را برای پردازش به‌وجود نمی‌آورند. برخی از حروف به دو شکل عربی و فارسی در توکن‌ها ظاهر می‌شوند. برای مثال دو «مجله» و «مجلة». بنابراین قبل از پردازش حروف عربی به حروف فارسی تبدیل و توکن‌ها اصلاح می‌شوند.

سه برچسب‌زن سی آر اِف برای استخراج فراداده‌های سرآیند، فراداده‌های مراجع فارسی و فراداده‌های مراجع انگلیسی استفاده می‌شود. فراداده‌های متنی براساس پیکره متون مقاله‌ها تعریف شده‌اند. برای هر برچسب‌زن، مجموعه‌ای از ویژگی‌ها و برچسب‌های مجزا تعریف شده است. مجموعه‌ای از ویژگی‌های تعریف شده برای برچسب‌زنی، از هر یک از توکن‌های مربوطه استخراج می‌شود. این مجموعه، بردار ویژگی[1] نام دارد. هر الگوریتم برچسب‌زنی سی آر اِف با دریافت مجموعه‌ای از بردارهای ویژگی و برچسب‌های متناظر، مدل یادگیری را ایجاد می‌کند. پس از یادگیری، داده‌های جدید با مدل ایجاد شده برچسب‌زنی می‌شوند.

فراداده‌های سرآیند شامل عنوان، نویسندگان، اطلاعات نویسندگان، چکیده، کلمات کلیدی، نام مجله، دوره، شماره و صفحه مجله و تاریخ است. این فراداده‌ها، برچسب‌های الگوریتم سی آر اِف سرآیند هستند. این برچسب‌زن دنباله‌ای از کلمات سرآیند را برچسب‌زنی می‌کند. ویژگی‌هایی که برای برچسب‌زنی سرآیند تعریف شده‌اند، به شکل زیر است:

- توکن
- یک توکن با رقم شروع شده یا پایان یافته باشد. بر این اساس این ویژگی سه مقدار به خود می‌گیرد.
- توکن شامل علامت نقطه باشد.
- کاراکترهای توکن، انگلیسی یا فارسی باشند.
- توکن شامل علامت نگارشی باشد.
- توکن با الگوی ایمیل تطبیق پیدا کند.
- توکن یکی از کلمات عنوان بخش کلمات کلیدی مانند «کلیدواژه» باشد.

1. Feature Vector

- توکن شامل کلمه «چکیده» باشد.
- توکن شامل یکی از کلماتی که اغلب در اطلاعات مجله به‌کار می‌رود، مانند «مجله» و «دوره» باشد.
- توکن شامل یکی از کلماتی که اغلب در اطلاعات نویسندگان به‌کار می‌رود، مانند «استادیار» و «دانشجو» باشد.
- توکن شامل کلمات ماه و فصل مانند «بهار» و «اردیبهشت» یا الگوی ارقام سال مانند «۱۳۹۳» باشد.
- توکن برابر الگوی صفحات مانند «ص.» و «۱۰۰-۱۱۰» باشد.
- توکن فقط شامل عدد باشد.
- توکن یکی از نام‌های شهرهای ایران باشد.
- توکن یکی از نام‌های ایرانی مردان و زنان باشد.
- شماره سطری که توکن در آن قرار گرفته است.
- موقعیت توکن در سطر. این ویژگی یکی از سه مقدار ابتدا، وسط و انتهای سطر را به خود می‌گیرد.

فراداده‌های مراجع فارسی شامل نویسنده، عنوان، تاریخ، منبع، دوره، شماره، مؤسسه، صفحه و مکان است. فراداده منبع در مراجع می‌تواند نام مجله، نام انتشارات، نام کنفرانس و پایان‌نامه باشد. در الگوریتم سی آر اِف مراجع فارسی، یک برچسب مشترک برای فراداده دوره و شماره استفاده شده است. برای هر یک از فراداده‌های دیگر، یک برچسب مجزا تعریف شده است. دنباله توکن‌های هر مرجع فارسی توسط در این الگوریتم برچسب‌زنی می‌شود. ویژگی‌های تعریف شده برای این سی آر اِف شامل موارد زیر است:

- توکن
- توکن شامل کلمات ماه و فصل مانند «بهار» و «اردیبهشت» یا الگوی ارقام سال مانند «۱۳۹۳» باشد.
- توکن برابر الگوی صفحات مانند «ص.» و «۱۰۰-۱۱۰» باشد.
- توکن شامل ارقام یا حروف یا هر دو باشد.
- توکن برابر الگوی مخفف اسامی مانند «م.،» باشد.
- توکن شامل علامت نگارشی باشد. این ویژگی شش مقدار را براساس نوع علامت نگارشی به خود می‌گیرد.
- توکن شامل کلماتی مانند «مجله» باشد که در فراداده منبع به‌کار می‌رود.
- توکن شامل کلماتی که در مؤسسه به‌کار می‌رود مانند «دانشگاه» باشد.

- توکن شامل کلماتی که در دوره و شماره مجله به‌کار می‌رود مانند «سال» باشد.
- توکن فقط شامل الفبای انگلیسی باشد.
- توکن برابر اسامی یکی از شهرهای ایران باشد.
- توکن شامل یکی از نام‌های مردان و زنان ایرانی باشد.
- موقعیت توکن در هر مرجع. برای این ویژگی پنج موقعیت مکانی براساس تعداد توکن‌های مرجع تعریف شده است.
- اولین توکن مرجع باشد.

فراداده‌های مراجع انگلیسی، عنوان، نویسنده، تاریخ، منبع، دوره، شماره، مکان و صفحه است. برای فراداده‌های دوره و شماره یک برچسب مشترک و برای هر یک از فراداده‌های دیگر، یک برچسب مجزا در الگوریتم سی آر اِف مراجع انگلیسی در نظر گرفته شده است. دنباله توکن‌های هر مرجع انگلیسی توسط این الگوریتم برچسب‌زنی می‌شود. ویژگی‌های زیر برای این برچسب‌زنی تعریف شده است:

- توکن
- الگوی ارقام سال مانند "2009" و کلمات ماه‌های میلادی مانند "May"
- الگوی صفحات مانند ".pp" و "page"
- شروع توکن با حروف بزرگ انگلیسی
- توکن شامل حروف، اعداد یا هر دو باشد.
- الگوی مخفف اسامی مانند ".C" و ".A.C"
- توکن شامل کلمات .et al باشد.
- توکن شامل علائم نگارشی باشد. براساس نوع علامت نگارشی این ویژگی شش مقدار مختلف را دریافت می‌کند.
- مجموعه کلماتی مانند "journal" که در فراداده منبع اغلب به‌کار می‌رود.
- مجموعه کلماتی مانند "volume" که در فراداده دوره و شماره اغلب به‌کار می‌رود.
- مکان توکن در مرجع. این ویژگی براساس تعداد توکن مرجع محاسبه می‌شود و پنج مقدار دارد.

پس از اینکه سه الگوریتم سی آر اِف، توکن‌های سرآیند، مراجع فارسی و مراجع انگلیسی را برچسب‌زنی کردند، توکن‌های برچسب خورده همسایه که برچسب یکسان دارند، به یکدیگر متصل می‌شوند و فراداده‌ها را تشکیل می‌دهند.

تفکیک دوره و شماره

توکن‌هایی که در برچسب‌زنی مراجع انگلیسی و فارسی برچسب دوره را دریافت کرده‌انـد، شـامل دو فراداده دوره و شماره هستند. پس از برچسب‌زنی و اتصال توکن‌های همسایه، ایـن دو فـراداده براسـاس الگوهای متنی، ارقام و علائم نگارشی شناسـایی مـی‌شـوند. در مراجع فارسـی، الگـوی «دوره (شـماره):»، کلمات «سال»، «دوره»، «جلد»، «شماره»، «پیاپی»، مخفف هر یـک از ایـن کلمـات ماننـد «س» و در مراجع انگلیسی الگوی "Volume (number)"، کلمات "Vol"، "No"، "Volume"، "Number" و"Issue" به-کار برده می‌شود.

آزمایش

برای تست روش پیشنهادی، ۱۰۰ مقاله مجلات مختلف علمی- پژوهشی ایران انتخاب شده اسـت. برای این کار مجموعه‌ای از مجلات علمی که در سامانه ارزیابی نشریات وزارت علوم ثبت شده‌اند استفاده شده است. مجموعه‌ای از مقاله‌ها به شکل تصادفی از این مجلات انتخاب شده‌اند. انتخاب مقاله‌ها به شکلی انجام شده است که تنوع فرمت در پیکره متون وجود داشته باشد. این مقاله‌ها در حوزه‌های علمی مختلـف هستند و به شکل رایگان در وب سایت مجلات در دسترس هستند. تعـداد مقالـه‌هـای انتخـاب شـده مشـابه پژوهش‌هایی در زبان انگلیسی مانند کواسویچ و همکارانش (۲۰۱۱) است که برای ارزیابی به شکل دسـتی پیکره متون را برچسب‌زنی کرده‌اند. مقاله‌های جمع‌آوری شده به فرمت متن تبدیل شده‌اند. سرآیند مقاله‌ها با فراداده‌های این بخش برچسب‌زنی شده‌اند. مجموعه‌ای از مراجع انگلیسـی و فارسـی از ایـن مقالـه‌هـا بـا فراداده‌های آنها برچسب‌زنی شده‌اند. در مجموع ۱۰۰ سرآیند و ۳۴۲ مرجع فارسی و ۴۰۰ مرجع انگلیسـی برای تست، برچسب‌زنی و استفاده شده‌اند.

تست با روش اعتبارسنجی عرضی پنج قسمتی[1] اجرا شده است. در ایـن‌روش پیکـره متـون بـه پـنج قسمت مساوی تقسیم و پنج بار تست انجام می‌شود. در هر بار تست، یکی از پنج بخش پیکـره متـون بـرای تست و چهار بخش دیگر برای یادگیری و ایجاد مدل در الگوریتم برچسب‌زنی استفاده می‌شود. در نهایت برای ارزیابی، میانگین نتایج پنج تست محاسبه می‌شود. برای پیاده‌سازی الگوریتم برچسب‌زنی از کتابخانـه سی آر اِف پلاس پلاس[2] استفاده شده است. تخمین پارامترها در این کتابخانه با الگوریتم L-BFGS انجام می‌شود.

معیارهای ارزیابی

1. five- fold cross validation
2. CRF++

برای ارزیابی از سه معیار دقت[1]، فراخوانی[2] و معیار اِف[3] استفاده شده است. این ارزیابی برای توکن‌های برچسب‌زده اندازه‌گیری شده است؛ بنابراین به شکل زیر معیارهای ارزیابی تعریف می‌شوند:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (۷) \qquad \text{Recall} = \frac{TP}{TP+FN} \qquad (۸)$$

$$\text{F-Measure} = \frac{2*Precision*Recall}{Precision+Recall} \qquad (۹)$$

در رابطه‌های بالا TP[4] تعداد توکن‌هایی است که برچسب آنها عنوان یا نویسنده یا فرادادەهای دیگر بوده و درست پیش‌بینی شده است. FP[5] تعداد توکن‌هایی نامربوط به داده است که برچسب آنها که اشتباه نسبت داده شده است. برای مثال هنگام بررسی عنوان، یک توکن جزء نویسنده است و به اشتباه برچسب عنوان را دریافت کرده است. FN[6] تعداد توکن‌های مربوط به داده است که به اشتباه برچسب دیگری را دریافت کرده‌اند. برای مثال هنگام ارزیابی عنوان، یک توکن که جزء توکن‌های عنوان بوده، به اشتباه برچسب چکیده را دریافت کرده است.

نتایج

نتایج به‌دست آمده برای فرادادەهای سرآیند در جدول ۱ نمایش داده شده است. دقت استخراج تمامی فرادادەها در سطح خوبی است. بیشترین دقت برای فراداده چکیده با معیار اِف برابر ۹۹/۶ درصد است. این فراداده تعداد توکن بسیار بیشتری نسبت به بقیه فرادادەها دارد. فراداده نویسنده با معیار اِف برابر ۹۴/۳ درصدکمترین دقت را دارد. نتایج به‌دست آمده برای فرادادەهای مراجع فارسی در جدول ۲ آمده است.

نتایج نشان می‌دهد دقت اکثر فرادادەها بالای ۹۰ درصد است. دقت فراداده مؤسسه با معیار اِف برابر ۸۰/۹۵ درصد کمتر از بقیه است. تعداد این فراداده در پیکره متون نسبت به فرادادەهای دیگر کمتر است. علاوه بر این کلمات نحوی که در این فراداده به‌کار می رود، تنوع بیشتری دارد. این دو عامل در پایین آمدن دقت مؤثر است. فراداده مکان بعد از مؤسسه دقت پایینی نسبت به فرادادەهای دیگر دارد.

1. Precision
2. Recall
3. F-measure
4. True positive
5. False positive
6. False negative

در مراجع فارسی اسامی شهرها در فرادادههای مکان و مؤسسه به‌کار می‌رود. این مسئله باعث می‌شود در برخی از موارد فرادادههای مکان و مؤسسه به اشتباه تشخیص داده شوند. بهترین نتیجه برای فراداده تاریخ با معیار اِف برابر ۹۸/۳۷ درصد به‌دست آمده است.

جدول ۱. نتایج برچسب‌زنی فرادادههای سرآیند مقاله

فراداده	دقت	فراخوان	معیار اِف
عنوان	۹۶/۵۸	۹۶/۸۲	۹۶/۶۷
نویسنده	۹۵/۷۳	۹۳/۱۸	۹۴/۳۰
اطلاعات نویسندگان	۹۷/۹۹	۹۷/۰۹	۹۷/۵۲
چکیده	۹۹/۵۲	۹۹/۶۹	۹۹/۶۰
کلمات کلیدی	۹۶/۴۹	۹۶/۵۴	۹۶/۴۲
مجله	۹۴/۲۹	۹۸/۶۶	۹۶/۱۳
دوره	۹۹/۵۳	۹۶/۹۵	۹۸/۲۰
شماره	۹۸/۰۹	۹۵/۵۹	۹۶/۷۷
صفحه	۹۷/۷۷	۹۴/۳۵	۹۵/۶۹
تاریخ	۹۷/۱۷	۹۸/۱۹	۹۷/۶۴
کل فرادادهها	۹۷/۳۱	۹۶/۷۰	۹۶/۸۹

جدول ۲. نتایج برچسب‌زنی فرادادههای مراجع فارسی مقاله

فراداده	دقت	فراخوان	معیار اِف
عنوان	۹۶/۳۳	۹۸/۸۰	۹۷/۵۲
نویسنده	۹۸/۹۶	۹۸/۵۳	۹۸/۷۲
تاریخ	۹۷/۹۵	۹۸/۸۳	۹۸/۳۷
منبع	۹۳/۳۸	۹۱/۲۹	۹۲/۲۳
دوره	۹۸/۴۷	۹۷/۱۸	۹۷/۷۹
صفحه	۹۷/۵۵	۹۸/۰۹	۹۷/۷۷
مؤسسه	۸۳/۳۴	۷۹/۱۴	۸۰/۹۵
مکان	۹۲/۰۷	۸۳/۸۶	۸۷/۶۲
کل فرادادهها	۹۴/۷۵	۹۳/۲۱	۹۳/۸۷

جدول ۳. نتایج برچسب‌زنی فرادادههای مراجع انگلیسی مقاله

فراداده	دقت	فراخوان	معیار اِف

عنوان	۹۶/۱۰	۹۸/۲۷	۹۷/۱۵
نویسنده	۹۷/۹۰	۹۹/۳۰	۹۸/۵۸
تاریخ	۹۸/۳۳	۹۶/۹۱	۹۷/۶۰
منبع	۹۳/۲۳	۹۲/۲۰	۹۲/۶۵
دوره	۹۷/۶۳	۹۸/۰۱	۹۷/۸۰
صفحه	۹۹/۳۱	۹۸/۷۹	۹۹/۰۳
مکان	۸۸/۱۵	۷۴/۸۲	۸۰/۴۹
کل فرادادهها	۹۵/۸۰	۹۴/۰۴	۹۴/۷۵

در زبان فارسی کلماتی که به شکل مشترک در فرادادههای مختلف بهکار میروند نسبت به زبان انگلیسی بیشتر است. برای مثال بسیاری از اسامی ایرانی که برای نام افراد بهکار میرود با معانی دیگر در فرادادههای دیگر استفاده میشود. این مسئله ممکن است باعث بروز خطا شود. اکثر خطاهای بهوجود آمده در استخراج فرادادهها مربوط به توکنهایی است که در مرز دو فراداده قرار دارند. نتایج استخراج فرادادههای مراجع انگلیسی در جدول ۳ آمده است. فراداده صفحه از دقت بالاتری برخوردار است. فراداده صفحه تنوع کمتری دارد.

جدول ۴. مقایسه نتایج برچسبزنی فرادادههای سرآیند مقاله

فراداده	این مقاله (سرآیند فارسی)	CRF (Peng and McCallum, 2006)	SVM (Han et al., 2003)
عنوان	۹۶/۶۷	۹۷/۱۰	۹۶/۵
نویسنده	۹۴/۳۰	۹۷/۵۰	۹۷/۲
اطلاعات نویسندگان	۹۷/۵۲	۹۷	۹۳/۸
چکیده	۹۹/۶۰	۹۹/۷	۹۳/۸
کلمات کلیدی	۹۶/۴۲	۸۸/۸	۸۸/۵
تاریخ	۹۷/۶۴	۹۵	۹۰/۲
کل فرادادهها	۹۷/۰۲	۹۵/۸۵	۹۳/۳

جدول ۵. مقایسه نتایج برچسبزنی فرادادههای مراجع مقاله

فراداده	این مقاله (مراجع فارسی)	این مقاله (مراجع انگلیسی)	CRF (Peng and McCallum, 2006)	CRF (Councill, Giles and Kan, 2008)
عنوان	۹۷/۵۲	۹۷/۱۵	۹۸/۳	۹۷
نویسنده	۹۸/۷۲	۹۸/۵۸	۹۹/۴	۹۹
تاریخ	۹۸/۳۷	۹۷/۶۰	۹۸/۹	۹۹
منبع	۹۲/۲۳	۹۲/۶۵	۹۱/۳	۹۱
دوره	۹۷/۷۹	۹۷/۸۰	۹۷/۸	۹۶

صفحه	۹۷/۷۷	۹۹/۰۳	۹۸/۶	۹۸
مؤسسه	۸۰/۹۵	–	۹۴	۸۹
مکان	۸۷/۶۲	۸۰/۴۹	۸۷/۲	۹۳
کل فرادادەها	۹۳/۸۷	۹۴/۷۵	۹۵/۶۸	۹۵/۲۵

نتایج روشی که در این مقاله ارائه شده است، با سه پژوهشی که در زبان انگلیسی انجام شده است، مقایسه شده است. جدول ۴ نتایج مقایسه فرادادەهای سرآیند و جدول ۵ نتایج مقایسه فرادادەهای مراجع را نشان می‌دهد. برای مقایسه از معیار اِف استفاده شده است. فرادادەهایی که در این مقاله و کارهای دیگر مشترک هستند، در نظر گرفته شده است. برای فراداده منبع در پژوهش‌های زبان انگلیسی، از نتایج فراداده مجله استفاده شده است.

مقایسه میانگین نتایج به‌دست آمده نشان می‌دهد در فرادادەهای سرآیند نتایج پژوهش این مقاله بهتر از عملکرد دو پژوهش انجام شده در زبان انگلیسی است. نتایج استخراج فراداده نویسنده در پژوهش‌های زبان انگلیسی بهتر است. برای فراداده چکیده در پژوهش‌های انگلیسی زبان، نتایج ضعیف‌تری به‌دست آمده است. مقایسه میانگین نتایج استخراج فرادادەهای مراجع، نشان می‌دهد پژوهش‌های زبان انگلیسی با اختلاف یک تا دو درصد بهتر است. نتایج استخراج فراداده مؤسسه نسبت به فرادادەهای دیگر ضعیف‌تر است. مقایسه نتایج به‌دست آمده حاکی از آن است که اختلاف نتایج در اکثر فرادادەها بین یک تا دو درصد است.

نتیجه

استخراج فرادادەهای متنی از مقالات علمی برای نمایه‌سازی مقالات لازم است. این کار به‌دلیل تنوع مقاله‌ها یک مسئله به‌شمار می‌آید. در این مقاله یک چارچوب برای استخراج فرادادەهای متنی از مقاله‌های علمی فارسی پیشنهاد شد. این چارچوب شامل شناسایی سرآیند و بخش مراجع، تفکیک مراجع و پردازش سرآیند و هر یک از مراجع است. فرادادەهای سرآیند و فرادادەهای مراجع فارسی و مراجع انگلیسی در این چارچوب استخراج شد. پردازش سرآیند و هر یک از مراجع توسط الگوریتم برچسب‌زنی سی آر اِف انجام گرفت.

به‌طور میانگین مقدار معیار اِف در سطح توکن برای فرادادەهای سرآیند ۹۶/۸۹ درصد، برای فرادادەهای مراجع فارسی ۹۳/۸۷ درصد و برای فرادادەهای مراجع انگلیسی ۹۴/۷۵ درصد به‌دست آمده است. نتایج به‌دست آمده نشان می‌دهد که الگوریتم برچسب‌زنی سی آر اِف برای استخراج فرادادەهای

متنی از مقاله‌های فارسی عملکرد خوبی دارد. افزایش تعداد مقاله‌ها در پیکره متون می‌تواند خطا را کاهش دهد.

نتایج این مقاله با نتایج سه پژوهش انجام شده در زبان انگلیسی مقایسه شده است. نتایج به‌دست آمده نشان می‌دهد که عملکرد پژوهش این مقاله در زبان فارسی و پژوهش‌های انگلیسی زبان در استخراج فراداده‌ها، به یکدیگر نزدیک است.

کتابنامه

Beel, J., Gipp, B., Shaker, A., & Friedrich, N. (2010). SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size). *Proceedings of the 14th European Conference on Digital Libraries.* Glasgow.

Candeias, R. (2011). Metadata Extraction from Scholarly Articles.

cb2bib overview. (2016). Retrieved 2015, from http://www.molspaces.com/d_cb2bib-overview.php.

Councill, I. G., Giles, C. L., & Kan, M. Y. (2008). ParsCit: an Open-source CRF Reference String Parsing Package. *In LREC, 8*, 661-667.

Giuffrida, G., Sheck, E., & Yang, J. (2000). KnowledgeBased Metadata Extraction from PostScript Files. *Proceedings of the fifth ACM conference on Digital libraries* (pp. 77-84). San Antonio, TX, USA: ACM.

Guo, Z., & Jin, H. (2011a). A Rule-based Framework of Metadata Extraction from Scientific Papers. *10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science* (pp. 400-404). Wuxi: IEEE.

Guo, Z., & Jin, H. (2011b). Reference Metadata Extraction from Scientific Papers. *12th International Conference on Parallel and Distributed Computing, Applications and Technologies* (pp. 45-49). Gwangju: IEEE.

Han, H., Giles, C. L., Manavoglu, E., Zha, H., Zhang, Z., & Fox, E. A. (2003). Automatic document metadata extraction using support vector machines. *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on* (pp. 37-48). IEEE.

Hetzner, E. (2008). A simple method for citation metadata extraction using hidden markov models. *In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* (pp. 280-284). ACM.

Kovacevic, A., Ivanovic, D., Milosavljevic, B., Konjovic, Z., & Surla, D. (2011). Automatic Extraction of Metadata from Scientific Publications for CRIS Systems. *Electronic Library and Information Systems, 45* (4), 376-396.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the eighteenth international conference on machine learning. 1*, pp. 282-289. USA: Morgan Kaufmann.

ParsCit: Anopen-source CRF Reference String and Logical Document Structure Parsing Package. (2016). Retrieved 2015, from http://aye.comp.nus.edu.sg/parsCit/.

Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. *Information processing & management, 42* (4), 963-979.

Seymore, K., McCallum, A., & Rosenfeld, R. (1999). Learning hidden Markov model structure for information extraction. *AAAI-99 Workshop on Machine Learning for Information Extraction.*

Tkaczyk, D., Szostek, P., Dendek, P., Fedoryszak, M., & Bolikowski, L. (2015). CERMINE: automatic extraction of structured metadata from scientific literature. *IJDAR, 18* (4), 317-335.

Wallach, H. (2004). *Conditional Random Fields: An Introduction.* University of Pennsylvania CIS Technical Report.

Zhang, X., Zou, J., Le, D., & Thoma, G. R. (2011). A structural SVM approach for reference parsing. *BMC bioinformatics*, *12* (3), 479-484.